<Bioinformatics Center>Bio-knowledge Engineering

Bioinformatics Center
– Bio-knowledge Engineering –

http://www.bic.kyoto-u.ac.jp/pathway/index.html

Prof
MAMITSUKA, Hiroshi
(D Sc)

Assist Prof
NGUYEN, Canh Hao
(D Knowledge Science)

Program-Specific Res
WIMALAWARNE, Kishan
(D Eng)

Program-Specific Res
SUN, Lu
(D Eng)

Students

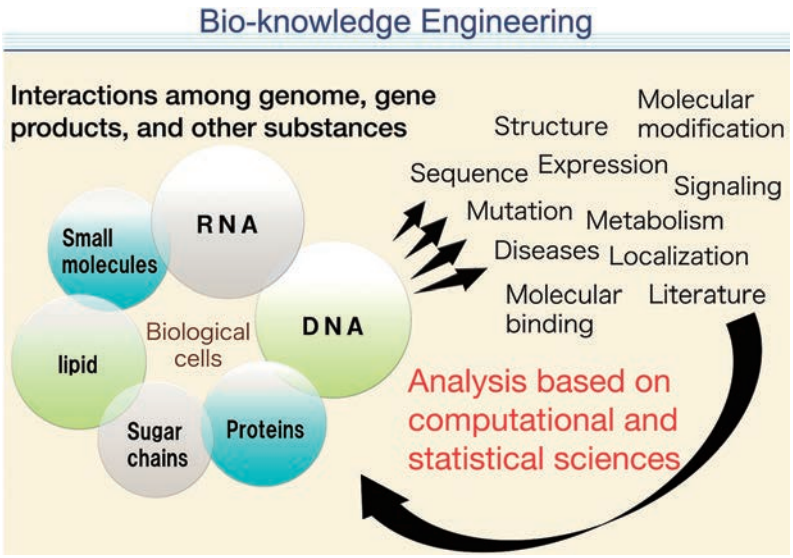
NGUYEN, Dai Hai (D1)
TOHZAKI, Yudai (M2)
KANEKO, Teruo (UG)

Guest Scholar

LI, Limin (Ph D) Xi'an Jiaotong University, China, P.R., 17 November–1 December

Scope of Research

We are interested in graphs and networks in biology, chemistry, and medical sciences, including metabolic networks, protein-protein interactions and chemical compounds. We have developed original techniques in machine learning and data mining for analyzing these graphs and networks, occasionally combining with table-format datasets, such as gene expression and chemical properties. We have applied the techniques developed to real data to demonstrate the performance of the methods and find new scientific insights.



KEYWORDS

Bioinformatics
Computational Genomics
Data Mining
Machine Learning
Systems Biology

Selected Publications

Yamada, M.; Lian, W.; Goyal, A.; Chen, J.; Wimalawarne, K.; Kahn, S.; Kaski, S.; Mamitsuka, H.; Chang, Y., Convex Factorization Machine for Toxicogenomics Prediction, *Proceedings of the Twenty-third ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)*, 1215-1224 (2017).

Karasuyama, M.; Mamitsuka, H., Adaptive Edge Weighting for Graph-Based Learning Algorithms, *Mach. Learn.*, **106(2),** 307-335 (2017).

Takigawa, I.; Mamitsuka, H., Generalized Sparse Learning of Linear Models over the Complete Subgraph Feature Set, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39(3),** 617-624 (2017).

Yotsukura, S.; Karasuyama, M.; Takigawa, I.; Mamitsuka, H., Exploring Phenotype Patterns of Breast Cancer within Somatic Mutations, *Brief. Bioinform.*, **18(4),** 619-633 (2017).

Yotsukura, S.; duVerle, D.; Hancock, T.; Natsume-Kitatani, Y.; Mamitsuka, H., Computational Recognition for Long Non-coding RNA (lncRNA): Software and Databases, *Brief. Bioinform.*, **18(1),** 9-27 (2017).

Distances on Graph with Global Information

Graph is an important topic in Machine Learning for many reasons. Graph, as a model of networks, is frequently used for network analysis, which found many applications in biological, chemical and social networks. Moreover, graph has been used in many other situation in analyzing complicated data. In high dimensional data, one of the most difficult situation for statistical analysis, neighborhood graph can encode data distribution to avoid the curse of dimensionality in many usual statistical models. Data on manifolds, which are too complicated for parametric models, are usually converted into graphs. Graph has surprising properties on other problem such as in semi-supervised learning, regularization and spectral clustering. It offers a nice way to estimate the number of clusters of data, which is usually a very difficult problem. In short, graph is a flexible tool to model data in difficult situations.

A graph encodes distribution of data, the common problem is that one needs to learn a statistical model from the graphs. As a graph usually represents similarity relationship among data points, it is the objective to learn a model that is smooth on the graph. Graph Laplacian is used to score how smooth a function is on graph. For its computational efficiency, most methods extract information on graph Laplacians to construct learning models, such as graph kernels, commute time, hitting time distance, resistance distances and Laplacian graph embedding.

However, it was recently discovered that in large graphs, graph Laplacians usually do not give meaningful information for learning problem. This is known as the global information loss problem. That is, spectral clustering would have almost random results. Semi-supervised learning models, usually extract information from unlabeled data, cannot benefit from large graphs encoding the data. It was later proven that for large enough graphs, graph Laplacians only contain local information, not suitable for learning models. This is a key problem with large data where graph Laplacians have been usually used. This result shows why graph fail to fulfill its potential of a flexible tool to model difficult distributions.

The aim of our research is to make graph a versatile tool for learning statistical models even for large graphs. We wish to estimate pairwise distances on a graph that reflect global information so that it can be used for learning models. Our idea is to solve the problem of resistance distance that an electrical flow on the graph would spread out too much, resulting in its energy function to be concentrated around its sources. We designed new resistance distances that account for this problem with a new energy function on the electrical flow. We also proved that the new distances keep global information. Our new distances were then applied to various problem of learning on graphs and show its usefulness for large data.

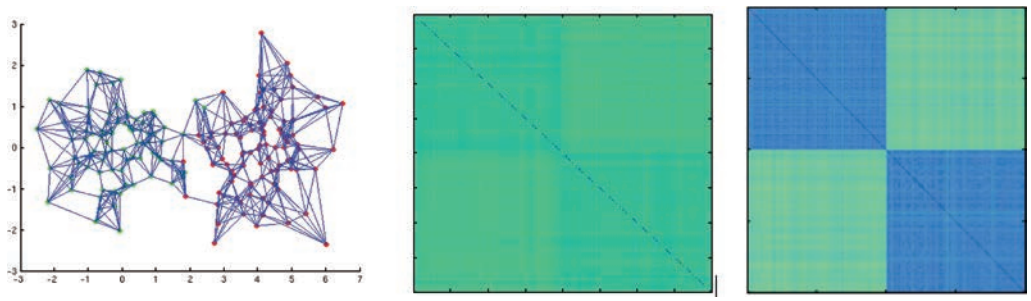


Figure 1. A graph (left) consists of two well-connected clusters. The distance based on graph Laplacian (heatmap in the middle heatmap) does not show cluster structure. Our new distance shows the two clusters by having two dark diagonal blocks on the heatmap (right).